METHOD FOR THE IDENTIFICATION OF CYTOSINE METHYLATION PATTERNS IN GENOMIC DNA

The invention concerns a method for the identification of 5-methyl-cytosine positions in genomic DNA.

The genetic information, which is obtained as a base sequence by complete sequencing of genomic DNA, only incompletely describes the genome of a cell. 5-Methylcytosine nucleobases, which are formed in the cell by reversible methylation of DNA, are an epigenetic information carrier and serve, for example, for the regulation of promoters. The methylation state of a genome represents the present status of gene expression, similar to an mRNA expression pattern.

5-Methylcytosine is the most frequent covalently modified base in the DNA of eukaryotic cells. It plays a role, for example, in the regulation of transcription, genomic imprinting and in tumorigenesis. The identification of 5-methylcytosine as a component of genetic information is thus of considerable interest. 5-Methylcytosine positions, however, cannot be identified by sequencing, since 5-methylcytosine has the same base pairing behavior as cytosine. Unfortunately, the epigenetic information that is carried by 5-methylcytosines becomes completely lost in PCR [polymerase chain reaction] amplification, and there is no method for obtaining this information by an amplification step.

Several methods are known, which solve these problems. For the most part, a chemical reaction or enzymatic treatment of the genomic DNA is

conducted, as a consequence of which, cytosine nucleobases can be distinguished from methylcytosine nucleobases. One current method is the reaction of genomic DNA with disulfite (also denoted bisulfite or pyrosulfite), which leads to the conversion of cytosine bases to uracil in two steps after alkaline hydrolysis (Shapiro, R., Cohen, B, Servis, R. Nature 227, 1047 (1970). 5-Methylcytosine remains unchanged under these conditions. The conversion of C to U leads to a change in the base sequence, from which the original 5-methylcytosines can now be determined by sequencing (only these [bases] will still supply a band in the C lane).

An overview of the other known possibilities for detecting 5-methylcytosines can be derived from the following review article together with the references belonging thereto: Rein, T., DePamphilis, M.L., Zorbas, H., Nucleic Acids Res. 26, 2255 (1998).

A method for characterizing specific DNA sequences is described in DD 293,139 A5, in which the DNA molecules, whose unmethylated recognition sites can be cleaved by an appropriate restriction endonuclease, are incubated in a reaction mixture with a second, unmethylated DNA species (particularly oligonucleotide duplexes, which contain the recognition site).

WO 97/46,705 A1 discloses a method for the detection of a methylated nucleic acid containing CpG, whereby the sample containing nucleic acid is brought into contact with a reagent, which modifies unmethylated cytosine, so that nucleic acids containing CpG in the sample are amplified by means of CpG-specific oligonucleotide primers, whereby the oligonucleotide primer

differentiates between modified methylated and unmethylated nucleic acids and detects methylated nucleic acids.

In addition, US 5,824,471 A1 describes a method for the determination of deviations between two nucleic acid strands, whereby a multiple number of duplexes are formed from the two strands or parts thereof and these duplexes are contacted with a first and a second different bacteriophage resolvase and whereby it is then established from which bacteriophage resolvase the duplex is cleaved, whereupon the differences are determined thereby.

However, it is not always necessary to actually determine the entire sequence of a gene or gene segment, as is the objective in the case of sequencing. This is particularly true if only a few 5-methylcytosine positions are to be found within a long base sequence in the case of a multiple number of different samples. Here, sequencing supplies essentially redundant information and is also very expensive. This is also true in the case when the sequence is already known and only the methylation positions need to be shown. It is also conceivable that in several cases in general, only the differences in the methylation pattern between different genomic DNA samples are of interest and that the determination of a multiple number of corresponding methylated positions as well as sequencing can be dispensed with. For the questions posed here, up until now, there has existed no method which supplies the desired result in a cost-favorable manner without sequencing each individual sample.

The sequence information per se is also continually less novel, since genome projects, whose goal is the complete sequence of various organisms,

are swiftly advancing. In fact, currently, even though only approximately 5% of the human genome has been completely sequenced, the study is progressing rapidly, since other genome projects have been completed and in this way sequencing resources have been freed up, so that every year another 5% is added. It is calculated that the sequencing of the human genome will be completed by the year 2006.

Matrix-assisted laser desorption/ionization mass spectrometry (MALDI) is a new, very high-performing development for the analysis of biomolecules (Karas, M. and Hillenkamp, F. 1988. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Anal. Chem. 60:2299-2301). An analyte molecule is embedded in a matrix absorbing in the UV. The matrix is evaporated in vacuum by a short laser pulse and the analyte is transported into the gas phase unfragmented. An applied voltage accelerates the ions in a field-free flight tube. Ions are accelerated to varying degrees based on their different masses. Smaller ions reach the detector sooner than larger ions. The time-of-flight is converted to the mass of the ions. Presently, this technology can distinguish molecules with a mass difference of 1 Da in the mass region from 1,000 to 4,000 Da. Due to the natural distribution of isotopes, most biomolecules, however, are approximated within 5 Da. Technically, this mass-spectrometric method can be very suitable for the analysis of biomolecules, but in order to distinguish them, the products that are to be analyzed must lie at least 5 Da apart from one another. Therefore, 600 molecules can be distinguished in this mass region. In the region between 4,000 and 100,000 Da, isotope

resolution is no longer achieved, but this region can also be used. Recently, the application of an infrared (IR) laser coupled with the MALDI analysis of DNA has been described (Berkenkamp, S., Kirpkar, F. and Hillenkamp, F. 1998. Infrared MALDI mass spectrometry of large nucleic acids. Science. 281: 260-262). It was possible by means of this combined technique to detect DNA fragments with a size of up to 2,500 bases.

Chemical mismatch cleavage is a method by means of which small differences between two single strands of DNA can be indicated (Cotton, R.G.H., Rodriguez, N.R. and Campbell, R.D. 1988. Reactivity of cytosine and thymine in single-base-pair mismatches with hydroxylamine and osmium tetroxide and its application to the study of mutations. Proc. Natl. Acad. Sci. USA. 85: 4397-4401; Cotton, R.G.H. 1993. Current methods for mutations detection. Mut. Res. 285: 125-144; Saleeba, J.A., and Cotton, R.G.H. 1993. Chemical cleavage of mismatch to detect mutations. Methods in Enzymology. 217: 286-295; Smooker, P.M. and Cotton, R.G.H. 1993. The use of chemical reagents in the detection of DNA mutations. Mutations Res. 288: 65-77). The chemical reactivity of C and T relative to osmium tetroxide and of C relative to hydroxylamine is increased, if these are not paired with their respective complementary bases. The nucleic acid strand is broken at the modified position by subsequent treatment with piperidine.

Another possibility to indicate non-complementary base pairs in heteroduplex DNA consists of the application of enzymes such as MutS, which bind to non-complementary base pairs (Smith, J. and Modrich, P. 1996.

Mutation detection with MutH, MutL, and MutS mismatch repair proteins. Proc. Natl. Acad. Sci. USA 93: 4374-4379; Parsons, B.L. and Heflich, R.H. 1997. Evaluation of MutS as a tool for direct measurement of point mutations in genomic DNA. Mut. Res. 374: 277-285).

At the present time, a rapid, cost-favorable and automatable method for finding methylated cytosines in genomic DNA is lacking. Such a method, however, is of great interest, since different methylation patterns can be drawn on in a variety of ways for characterizing cell types and thus can be used for diagnosis and classification of diseases (such as, for example, tumors) and this method could also be utilized, for example, for studies of cell differentiation.

The object of the present invention is thus to create a method for a cost-favorable parallelly-conducted dectection of epigenetic information carriers in the form of 5-methylcytosine bases in genomic DNA.

The object is resolved according to the invention by a method for the identification of 5-methylcytosine positions in genomic DNA, whereby the following method steps are conducted:

a) the genomic DNA of a cell, a cell line, a tissue or an individual is chemically treated in such a way that cytosine and 5-methylcytosine react differently and a different base-pairing behavior results for the two products in the duplex,

b) the same nucleic-acid segment is amplified by means of a polymerase reaction,

c) the same nucleic-acid segment of at least one other cell, cell line, tissue or individual or any desired reference DNA is treated according to steps a) and b),

d) heteroduplexes are formed from the at-least two amplified products of steps b) and c),

e) a detectable labeling is introduced into the heteroduplex by means of a reaction, which is specific to non-complementary base pairs.

According to the invention, it is preferred that for the identification of differences in the cytosine methylation pattern between various cells, cell lines, tissues and individuals, only those positions are applied and indicated, in which the cytosine methylation is variable between different cells, cell lines, tissues or individuals.

It is also preferred that a disulfite (bisulfite, pyrosulfite) is utilized as the reagent for the selective conversion of cytosine to uracil in step a), whereby 5-methylcytosine remains unchanged.

It is also preferred that genomic DNA of several individuals, tissues, cell lines or cells is jointly amplified in step b).

In addition, it is preferred that genomic DNA of several individuals, tissues, cell lines or cells is separately amplified and then jointly treated according to step e).

It is also preferred according to the invention that erroneous base pairings are produced by formation of heteroduplexes from the DNA of different individuals, tissues, cell lines or cells at the positions at which a 5-methylcytosine was localized in the genomic DNA.

It is also preferred that erroneous base pairings occur in step d) at those positions at which cytosine was found in the genomic DNA by the formation of heteroduplexes with a completely methylated reference DNA.

It is also preferred that erroneous base pairings occur in step d) at the positions at which 5-methylcytosine was found in the genomic DNA, by formation of heteroduplexes with a completely demethylated reference DNA.

In addition, according to the invention, it is preferred that the erroneous base pairings lead to a specific or sufficiently selective backbone cleavage at these positions by means of "chemical mismatch cleavage" (chemical modification at non-complementary positions).

It is also preferred that the DNA at the erroneous base pairings is cleaved enzymatically specifically or sufficiently selectively.

In the method according to the invention, it is also preferred that 1 DNA fragment is obtained in step e), the size of which provides an inference to the cleavage positions and thus to the position of the methylcytosines and/or the variable methylation positions between different individuals, tissues, cell lines or cells.

It is preferred that the analysis of size (molecular weights) of the DNA fragments is conducted by means of mass spectrometry.

It is particularly preferred that the fragments are analyzed by means of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI).

It is also particularly preferred that the fragments are analyzed by means of electrospray ionization mass spectrometry (ESI).

It is particularly preferred that the size of the fragments produced in step e) is adapted to the performance capacity of the mass spectrometer.

It is most particularly preferred that several PCRs of a gene segment are conducted and the primers are set stepwise such that the fragment size to be expected at least in one of these PCRs falls in the mass range that can be detected by means of mass spectrometry.

It is particularly preferred that one of the PCR primers is newly positioned stepwise by the maximally detectable mass range of the mass spectrometer relative to the other primer.

It is preferred according to the invention that in step b) one primer of the PCR is provided with a chemical function, so that the PCR product can be immobilized on a surface.

It is also preferred according to the invention that the PCR product produced in step b) is transferred into different reaction vessels and the surfaces of the reaction vessels are treated chemically in such a way that the PCR product can be bound thereon.

It is also particularly preferred that PCR products of different individuals that are prepared in step c) are transferred into different reaction vessels prepared as described above.

In addition, it is preferred according to the invention that an enzyme is used for step e), which [enzyme] forms a complex with a non-complementary base pair.

It is very particularly preferred that this enzyme is Muts.

In addition, it is preferred that the enzyme bears a label, by means of which a complex can be visualized.

It is also preferred according to the invention that the label is a fluorescence label, a chemiluminescence label, a mass label or a photochemically cleavable mass label.

In addition, it is preferred according to the invention that an amplified DNA sample according to step c) in claim 1, which displays a difference relative to an amplified DNA sample in step b) is compared in a second run of the method itself [with] a [similar] DNA sample according to step b) in claim 1 and with all other DNA samples to be investigated.

It is also preferred according to the invention that a preselection of the gene segments to be investigated in detail by mass spectrometry will be conducted by means of a fluorescence labeling or chemiluminescence labeling of the immobilized DNA strand, the lack of which indicates the presence of methylated cytosines in the investigated genomic DNA segment after conducting steps d) and e) of claim 1 and a washing step.

It is further preferred according to the invention that a preselection of the gene segments to be investigated in detail by mass spectrometry is conducted by means of a more nonspecific variant according to claims 20 to 23.

Another subject of the present invention is a kit for conducting the method according to the invention, comprising DNA of at least two individuals, tissues, cell lines or cells that are as different as possible as well as reagents, in order to indicate the variable methylation positions.

The [kit] according to the invention also comprises completely methylated and/or demethylated DNA and reagents, which are necessary for the detection of methylated cytosines in any DNA sample.

The method according to the inventions serves for the identification of 5-methylcytosine positions in genomic DNA, which can be of the most varied origin. The genomic DNA is first treated chemically in such a way that a difference is produced in the reaction of cytosine bases and of to 5-methylcytosine bases. Possible reagents here include, e.g., disulfite (also denoted bisulfite or pyrosulfite), hydrazine and permanganate. In a preferred variant of the method, the genomic DNA is treated with disulfite in the presence of hydroquinone or hydroquinone derivatives, whereby the cytosine bases are converted to uracil selectively after subsequent alkaline hydrolysis. 5-Methylcytosine remains unchanged under these conditions. After a purification process, which serves for separating the excess disulfite, a specific segment of the pretreated genomic DNA is now amplified in a polymerase reaction. In a preferred variant of the method, the polymerase chain reaction is used here. Then, the same segment of another genomic DNA sample is amplified to the same extent. The two amplified products are combined, whereby heteroduplexes are partially formed. In a preferred variant of the method, this is

produced in such a way that one of the PCR primers bears a function suitable for immobilization and that only one strand of the amplified product of the first sample is immobilized and then a hybridization is performed with the amplified product of the second sample. In another preferred variant, a multiple number of different amplified products of the same nucleic-acid segment are hybridized in this manner with the immobilized single strand of the amplified product of the first sample, which was first distributed in many wells of a microtiter plate. Now one hybridization experiment can be conducted in each well.

After the hybridization, a method is conducted, which leaves behind a detectable label at those positions in which an erroneous base pairing occurs in the heteroduplex. In a preferred variant of the method, this is conducted by chemical mismatch cleavage, which leads to a break of the backbone at the positions where an erroneous base pairing has occurred. The fragments obtained in this way can be analyzed by any method that can indicate the size of DNA fragments. Such a method should ideally permit conclusions on any position in the amplified nucleic-acid segment of the sample, at which an erroneous pairing has occurred in the heteroduplex. Erroneous base pairings in the heteroduplex are present particularly if, in the DNA of one sample, cytosine was present at this position, which was converted to uracil, but in the other sample, 5-methylcytosine was present, which remained unchanged in the chemical pretreatment. The method can be utilized also for the comparison of two or more genomic DNA samples; in this case, the analysis of the fragments supplies only the differences in the methylation pattern between the two samples

in the respective amplified nucleic-acid segment. However, it is also possible to utilize a DNA as a reference which has been completely methylated or demethylated enzymatically at C. In this case, the analysis of the fragments supplies all 5-methylcytosine positions in the respective amplified nucleic-acid segment.

In a particularly preferred variant of the method, mass spectrometry is applied to the analysis of fragments. After a preliminary purification, the fragments can be analyzed in the MALDI mass spectrometer. Alternatively, the solutions can be analyzed by electrospray ionization mass spectrometry (ESI). It may be necessary to investigate the nucleic-acid segment in question in several substeps by newly positioning a primer stepwise in several PCRs, each time depending on the performance capability of the method and the instrument utilized, and thus various amplified products of the substeps result ("primer walking").

In a variant of the method, the erroneous base pairings – alternatively to the analysis of fragments after a backbone cleavage in the heteroduples – may also be detected by means of an enzyme, which forms a complex with a non-complementary base pair. In a preferred variant, this enzyme is MutS, which bears a label, e.g., a fluorescence, chemiluminescence or mass label.

In another variant of the method, the presence of erroneous base pairings, i.e., in this case also the presence of relevant information in the amplified nucleic-acid segment, is detected by a fluorescence or chemiluminescence labeling. In a preferred variant of the method, an

immobilized DNA strand of the amplified product of sample 1 is provided with a fluorescence label on the end that does not serve for the immobilization. Heteroduplexes are formed with the amplified product of sample 2, and these are subjected to a chemical mismatch cleavage. If a backbone cleavage occurs at the immobilized strand, then the fluorescence label disappears after a denaturing washing step, and if the strand is not cleaved, then the label remains. Only the amplified products that have been cleaved are subsequently investigated in more detail by mass spectrometry.

<u>Examples</u>

Example 1

Method for finding all methylated cytosine positions

The genomic DNA to be investigated that derives from a cell line, or as much as possible from only one cell, is divided into two reaction vessels and one-half of this is either completely methylated or demethylated enzymatically at the cytosine. The enzyme is thermally inactivated and then both parts are again combined and treated with disulfite and then alkali. After a purification, amplification is conducted by means of PCR.

The chemical mismatch cleavage that is specific for the C mismatch and that is now conducted leads to a cleavage at the positions of a corresponding heteroduplex, at which an originally methylated C was found, if a complete demethylation of one-half of the genomic sample was conducted. On the other hand, a cleavage occurs at all originally non-methylated positions, if a complete

methylation of one-half of the genomic sample was previously conducted. For reliability, both methylation as well as demethylation may be conducted as a reference, but in this case, these procedures must be performed separately in two PCRs.

Variant 1: The above method is conducted in such a way that a primer is introduced in the PCR, which is functionalized such that a simple and specific immobilization is made possible after PCR. The immobilization is conducted onto beads or onto the surface of a microtiter plate. This permits the simple separation of components of the polymerase and mismatch cleavage reactions. After the chemical mismatch cleavage reaction, the duplex is thermally denatured and the solution is pipetted off. The DNA fragments from this solution are introduced onto a reversed-phase material and purified.

In the mass spectrometer, the fragments produce a "ladder" of peaks from which the methylated positions can be inferred. Theoretically, two peaks per CpG always occur at CpG postions on the basis of the symmetrical methylation; these peaks originate from the sense and anti-sense strands.

Variant 2: The reactions are conducted in solution and a purification is conducted after the individual reaction steps, if necessary, each time by means of a reversed-phase material.

Variant 3: Several individuals or cell types are parallelly investigated. A reference DNA is completely demethylated and then treated with disulfite. It is amplified by means of PCR after purification. A primer is again used, which bears a function suitable for the immobilization. The solution is distributed onto

the wells of a microtiter plate and immobilized. Then a hybridization is conducted against the PCR products from the samples also treated with disulfite, each time, one [product] per well (see also the described example with 97 individuals).

Variant 4: If the mass spectrometer cannot cover the [entire] measurement range, which would be necessary for the analysis of the total PCR product to determine methylations, the region of interest can also be found stepwise by conducting several PCRs and each time placing one of the primers closer to the other by the respective measurement range of the mass spectrometer. Thus, for example, only that region is detected, which lies between the primer to be shifted of the PCR in question and the next PCR. The method can be combined with the other variants.

Example 2

Method for finding positions with variable cytosine methylation

DNA of various individuals or cell lines is pooled and a treatment with disulfite is conducted as described above. After alkaline hydrolysis of the bisulfite adducts and purification of the product DNA, the latter is amplified by means of PCR. It is then purified again and after several minutes of reannealing at 25°C with OsO_4, the PCR product is cleaved at the positions with a C mismatch (chemical mismatch cleavage). A C → A mismatch then always occurs, if a methylated cytosine has been present only in several individuals prior

to the bisulfite treatment. In this process, possible SNPs (single nucleotide polymorphisms), as it were, also lead to the cleavage of the DNA. The latter must be distinguished from the methylated positions to be found, which is assured by employing the above-described method for finding all methylated cytosines.

The DNA product is now investigated by mass spectometry, as described above. If the initially generated PCR product is longer than can be detected with the currently available technology relative to mass spectrometry, then it is possible that the fragments produced by the chemical mismatch cleavage cannot be detected. In order to get around this, several PCRs can be conducted iteratively, i.e., one primer will always be kept constant, while the other primer will be positioned closer to the other primer continually in several steps, each time by the detection limit of the mass spectrometer (primer walking).

Example 3

Method with 97 individuals

A genomic segment of an individual (reference individual) is treated with disulfite and in this way, the cytosines are converted into uracils after subsequent alkaline hydrolysis of the bisulfite adduct. The methylcytosines remain unaffected in this reaction sequence. The product is purified and amplified by means of PCR. One of the PCR primers is provided on the 5' end with a chemical modification, which serves for immobilization. The product of this PCR is placed in the 96 wells of a microtiter plate and the PCR products are induced

to bind to the surface. Since only one primer is provided with the chemical modification for such binding, only one DNA strand binds to the surface. The plate is washed to eliminate the reagents of the binding chemistry and the complementary strands. In this way, the plate containing the reference DNA piece is prepared. The same genomic segment in each of the 96 other individuals is treated analagously with disulfite and then amplified. Each time two normal unmodified primers of the same sequence as for the reference individual are used for this PCR. The 96 PCR products are placed in the 96 wells of the prepared plate. By heating and slow cooling, the complementary strands of the 96 individuals are hybridized to the reference DNA (formation of the heteroduplex). The 96 individuals and reagents of the previous reactions [are] to be eliminated. An OsO_4 solution is added to each of the 96 wells, incubated, and then a backbone cleavage is induced with piperidine in a heteroduplex with a non-complementary base pair, one base of which is C. This will happen only if a methylcytosine is present instead of a cytosine in one strand of the heteroduplex, i.e. in one of the individuals. In this case, only the cytosine of one individual was converted to a uracil prior to the PCR, whereby a mismatch results in the heteroduplex with the counter-strand of another individual. The assay thus does not directly produce all methylated cytosines of a genomic segment, but only those that are variable between different individuals, tissues, cell lines or individual cells.

The heteroduplex is melted apart by heating and the solution is transferred to a mass spectrometer.

Example 4

Transfer of the solution to the mass spectrometer

One good variant is to take up the solution after melting the heteroduplex in a pipette tip, which is furnished with a reversed-phase material. The DNA products bind thereon by forming hydrophobic interactions by means of their trialkylammonium counter-ions and can thus be purified of the reagents of the chemical mismatch cleavage in several washing steps. The DNA products can be dissolved again from the reversed-phase material with 30% acetonitrile. This makes it possible to obtain the products directly on a prepared matrix on a MALDI target. After a single drying, the target is introduced into the mass spectrometer and the products [are] analyzed.

Example 5

Preselection by means of fluorescence labeling of gene segments relevant for methylation detection.

The genomic DNA to be investigated is immobilized on beads or an appropriately coated microtiter plate after the bisulfite reaction with subsequent PCR amplification as described above, in which one of the primers again bears a function that can serve for the subsequent immobilization. Completely demethylated DNA, treated like the sample DNA, is used as reference DNA and forms a heteroduplex with the immobilized sample DNA. Then an individual fluorescence-labeled base is attached enzymatically, for example, with terminal transferase to the 3' ends of the product. The subsequently conducted

"chemical mismatch cleavage" reaction in the case where a C/A mismatch is found in the product, leads to a cleavage of the immobilized strand, so that all fluorescence labelings are removed in the subsequent washing step after the thermal dehybridization. Thus a methylation was present within the amplified product, so that fluorescence no longer occurs in the microtiter plate or on the bead. The fluorescence only remains, if no mismatch is present, and thus no 5-methylcytosines are present in the gene segment in question. In the method, it must be considered that, e.g., SNPs may yield a false positive signal.

Accordingly, this method may be utilized also for simple, fluorescence-based detection of 5-methylcytosines in small gene segments, e.g. promoters. However, information may only be found of whether or not methylations are present in the region in question, but not how many and at which positions. Of course, this is compensated by a relatively small experimental expenditure and a good capability for conducting parallel experiments.

Then, analagously to one of the above-described examples, the precise position of methylcytosines can be determined by mass spectrometry in the gene segments classified as relevant.

Example 6

Preselection of heteroduplexes with mismatches

The heteroduplexes immobilized in one microtiter plate are first combined with a solution of MutS, to which a fluorescent dye is bound. Only the vessels, in which MutS has attached to erroneous base pairing positions, which is indicated

by the fact that the fluorescence can still be detected after several washing steps, are subsequently subjected to the chemical mismatch cleavage and analyzed in the mass spectrometer. In this way, time in the mass spectrometer and costs for purification are spared, since analysis of samples without detectable epigenetic information is avoided.